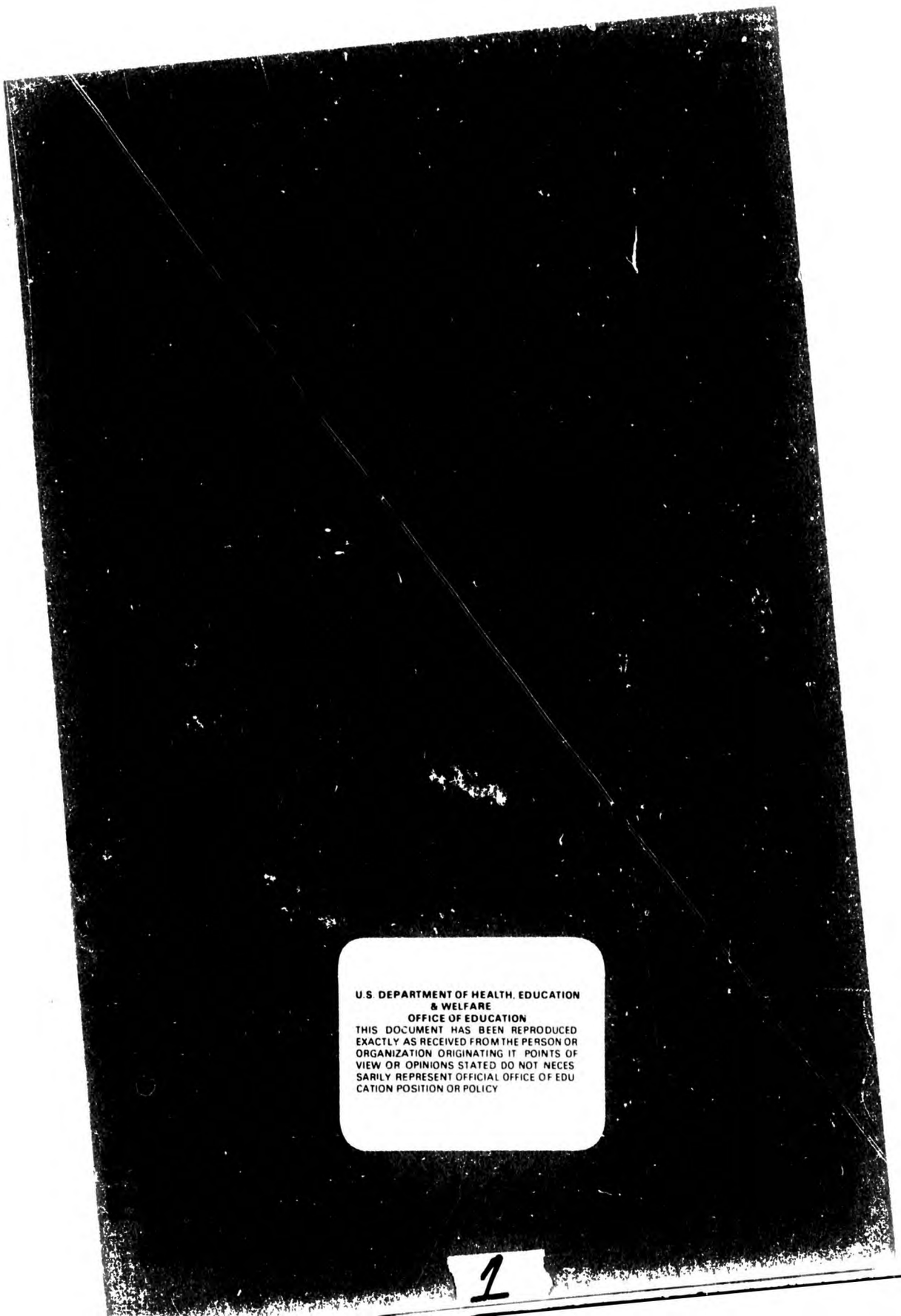
DOCUMENT RESUME

ED 051 294 TM 000 612

AUTHOR	Toole, Patrick F.; And Others
TITLE	Educational Quality Assessment Phase II Findings: Reliability and Validity.
INSTITUTION	Pennsylvania State Dept. of Education, Harrisburg.
PUB DATE	Aug 70
NOTE	14p.; Section 3
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Educational Quality, *Measurement Instruments, Questionnaires, Reliability, *State Surveys, Statistics, *Test Reliability, *Test Validity, Validity
IDENTIFIERS	*Pennsylvania Plan

ABSTRACT

Section 3 of Phase II of the Pennsylvania Plan is concerned with the adequacy of the educational quality assessment instruments. An overall discussion of reliability--content, criterion related, and construct--is presented. Reliability coefficients for the assessment inventories are provided and empirical studies of validity are described. (PR)



Phase II Findings

Section 3

Reliability and Validity

by Patrick F. Toole
Bureau of Educational Quality Assessment
Paul B. Campbell, Director
Joan S. Beers, Research Associate
Office of Research and Statistics

Pennsylvania Department of Education
August 1970

Commonwealth of Pennsylvania
Raymond P. Shafer, Governor

Department of Education
David H. Kurtzman, Secretary

Commissioner for Basic Education
B. Anton Hess

Assistant Commissioner for
Programs and Services
Donald M. Carroll Jr.

Bureau of Educational Quality Assessment
Thomas E. Kendig, Director (Acting)

Pennsylvania Department of Education
Box 911
Harrisburg, Pa. 17126

Contents

Foreword

The Pennsylvania Department of Education has been developing inventories in the affective areas to assess the quality of education in the public schools of the Commonwealth. Reliability and validity information indicate the adequacy of any measuring device. In the case of self-report inventories, however, validity aspects in particular are most difficult to determine.

Section 3 of *Phase II Findings* presents an overall discussion of reliability and validity, the reliability coefficients of the assessment inventories, and validity information collected about the inventories.

Reliability and Validity

> Reliability refers to consistency, to obtaining the same results again. Validity tells us whether the question or item really measures what it is supposed to measure. For instance, a clock is supposed to measure 'true' time, and to do so continuously. If it were to show the wrong time, we would say that it was invalid. If it were sometimes slow and sometimes fast, we would call it unreliable. It is possible to have a measure that is highly reliable yet of poor validity, for instance a clock that is precisely eighteen minutes fast consistently. The degree of reliability (consistency) sets limits to the degree of validity possible: validity cannot rise above a certain point if the measure is inconsistent to some degree. On the other hand, if we find that a measure has excellent validity, then it must also be reliable [Oppenheim, 1966, pp. 69-70].

As Oppenheim's clock analogy emphasizes, reliability and validity are interrelated concepts. Since reliability places limits on validity, how does one maximize it, thereby enhancing validity's possibilities? Fortunately, as Kerlinger (1964) notes, "Achieving reliability is to a large extent a technical matter [p. 459]." Lest this sound too optimistic a note, he quickly adds, "Validity, however, is much more than technique. It bores into the essence of science itself . . . [and] into philosophy [p. 459]." By science and philosophy, Kerlinger meant "truth" and the way we come to know the truth.

Reliability

Reliability studies may include one or more of the following techniques:

1. Split-halves coefficients computed in tests of sufficient length where scores from one-half of a test (sometimes the odd-numbered items) are correlated with scores from the other half (the even-numbered items) on a sample of the same subjects. The correlation is then expanded to compensate for the reduced range of the half test.

2. Test-retest coefficients calculated between the results of the same test taken by a single sample of respondents at two different times, several weeks or a month or two apart.

3. Internal consistency coefficients can be computed in a number of ways. Split-halves is one special case, a second is based upon the item to total score correlation, and a third case is based upon analysis of variance.

Internal consistency reliability coefficients were computed for all inventories in the educational quality assessment (EQA) battery. Where data were analogous to the usual right-wrong response form, the Kuder-Richardson

formula 20 was used. Where the response pattern was a matter of degree, the more general coefficient alpha (Cronbach, 1951) was computed. Duplicate study results from entirely different populations are available for some of the EQA inventories in Crites (1969) and Coopersmith (1967).

Perhaps another way of conceptualizing the internal consistency approach to reliability studies, whether it be based on an item to total score correlation or an analysis of variance approach, is to consider the entire inventory being studied as a time sequence and the responses to the separate items a series of tasks at separate consecutive times. Coefficients of internal consistency, based upon a suitable sample of subjects, identify those tasks which are consistent with each other in differentiating between high and low scoring subjects.

In most cases, each section of the EQA battery went through at least three stages of reliability analyses, with coefficients recomputed at each stage. Items were reworded, deleted or added according to the results of the previous stage. In some cases, an entirely different approach was necessary. Item decisions were based upon item to total score correlations, proportions of students omitting an item, discrimination indices, and factor analysis solutions. Table 1 presents the reliability coefficients after final refinement.

TABLE 1

Reliability Coefficients for EQA Inventories

	Inventory	Reliability	
		Grade 5 r_{tt}	Grade 11 r_{tt}
I	Self-Understanding	.87	.90
II	Understanding Others	.77	.88
III	Basic Skills	.90+[a]	.90+[b]
IV	Interest in School	.75	.85
V	Citizenship	.90	.91
VI	Health Habits	.82	.91
VII	Creative Potential	.82	.78
VII	Creative Output	—[c]	.93
VIII	Vocational Development	.77	.89
IX	Appreciation of Human Accomplishments	.79	.92
X	Preparation for Change	.79	.81

[a] Measured by the Stanford Achievement Battery or the Iowa Test of Basic Skills.

[b] Measured by the Stanford Achievement Battery or the Iowa Tests of Educational Development.

[c] Not measured.

Validity

One or more, preferably more, of three commonly accepted methods may be used to determine an inventory's validity. Hierarchically intertwined, they are:

Content Validity

Test items "must reflect the test's purposes, i.e., a social studies test consisting of factual content cannot be used validly to test hypotheses based on the theory of understanding or applying social studies ideas [Kerlinger, 1964, p. 447]." Techniques often used to assess content validity are observational analysis and content analysis, generally undertaken by a panel of independent experts. Here each item is judged individually in relation to the rationale, both theoretical and empirical, upon which the test or inventory is based.

Criterion Related Validity

A concurrent criterion, an independent measure of the same variable, is obtained and compared with the test's results to determine criterion related validity. Here, also, two or more independent criteria are better indicators of test validity than any one criterion. To the extent that the test and independent measures of the same attribute coincide, one may say the test has criterion related validity. But, as Oppenheim (1966) observes, even "If we happen to find pragmatic validity in respect to a particular criterion, we still need to know why it works, in terms of constructs [p. 75]."

Construct Validity

A construct can be defined "by a network of relations, all of which are anchored to observables and are so testable [Cronbach, 1969]." Such constructs, once obtained, would be expected to enter into relationships with other variables in predictable ways. Validity is inferred from such a predicted network of relationships; this validates both the measure and the theory behind it [Oppenheim, 1966]." Kerlinger (1964) says that "factor analysis may almost be called the most important of construct validity tools [p. 454]."

Some Opposing Views on Validity

Not surprising is the debate over and the differences among informed opinions on the concept of validity. Kerlinger's statement that validity "is concerned with the nature of 'reality'" itself implies questioning of scientific and/or philosophical truths.

Crites (1969) labels the concept of validity "ambiguous." He goes on to say that "It can be argued . . . that only the concept of 'content' validity

applies [to his scale]. Its relationships to empirically defined variables are neither instances of its validity or invalidity, although they are support for its significance or usefulness as a psychological concept [p. 50]."

Kerlinger (1964) quotes Rokeach as saying "that he was mainly preoccupied with construct validity," while Kerlinger himself notes that "any type of validation is construct validation [p. 2]."

Cronbach (1969) argues that "Validation is the task of the test interpreter. Others can do no more than offer him material to incorporate into his thinking . . . How one is to validate depends not on the test but on one's purpose in using the test. Since virtually no test is confined to a single purpose, it is illogical to speak of test validity. What one has to validate is a proposed interpretation of the test; for any test some interpretations are reasonably valid and others are not [p. 2]."

Given these caveats with respect to validity, we shall attempt to describe the efforts to validate several of the educational quality assessment inventories.

All of the EQA inventories have content validity. Their contents were "validated" according to the rationales developed for each of the ten goals* by the EQA staff or measurement researchers both locally and nationally. Many of the inventories were subjected to criterion related validity tests. Factor analyses from a sample of 3000 student responses for construct validity purposes are reported separately.*

Empirical Studies of Validity

Several studies were undertaken to assess validity under varying situations. In one study pupils with diverse reading abilities were administered three self-report inventories: Understanding Others, Citizenship and Appreciation of Human Accomplishments.

The study attempted to ascertain the effects of two different administrative procedures on pupils identified by teachers as being in the "lowest reading ability" groups.

To study the effects of two different administrative procedures, half of the *Ss* in two schools were randomly assigned to a read group and half to a nonread group. Treatment in the read group consisted of the test administrator reading each of the items to the *Ss*. The *Ss* responded in writing to each item after it was read. The nonread group read the items themselves.

Tables 2 and 3 illustrate the results between the read and nonread treatment groups.

---

*See Section 4 of *Phase II Findings*.

TABLE 2

Comparisons Of Means Between The
Read And The Nonread Treatment Groups

Inventory	School A results							
	Read group				Nonread group			
	N	$\bar{X}$	s	s^2	N	$\bar{X}$	s	s^2
Understanding Others	15	21.2	5.44	29.62	15	21.6	6.38	40.76
Citizenship	15	147.2*	18.0	326.3	15	160.7	15.9	255.8
Human Accomplishments	14	90.0**	10.5	110.8	15	86.4	19.2	368.9

*$t = -8.176$ $p < .01$
**$t = 2.264$ $p < .05$

TABLE 3

Comparisons Of Means Between The
Read And The Nonread Treatment Groups

Inventory	School B results							
	Read group				Nonread group			
	N	$\bar{X}$	s	s^2	N	$\bar{X}$	s	s^2
Understanding Others	15	24.6	4.8	23.9	15	25.2	3.9	15.3
Citizenship	15	172.3	16.7	280.4	15	171.8	15.3	234.2
Human Accomplishments	15	89.8*	11.4	131.6	15	86.0	9.6	93.5

*$t = 3.574$ $p < .01$

Significant differences between the read and the nonread means were found on the Citizenship inventory in school A and the Appreciation of Human Accomplishments inventory in schools A and B.

An interesting phenomenon occurred in school A with respect to the Citizenship inventory. Some of the items refer to honesty in unobserved situations, such as when one finds money or accidently "scratches a neighbor's car." When such items were read aloud by the test administrator, nervous laughter spread across the room. The effect was to reinforce the less socially desirable responses among the pupils, resulting in a significantly lower mean

score. In school B, the laughter did not occur and the means were not significantly different. The implications are that factors other than reading difficulty may have contributed to the differences between means in school A on the Citizenship inventory.

The significant differences between the read and nonread means on the Appreciation of Human Accomplishments inventory in both schools A and B imply that reading difficulty may have been a contributing factor. As a result of the study, the inventory was revised for 5th grade pupils.

In a second study, it was anticipated that *Ss* in urban and suburban high schools would score higher than *Ss* in a rural high school on three inventories: Understanding Others, Citizenship and Appreciation of Human Accomplishments.

Tables 4, 5 and 6 illustrate the results among the three inventories.

TABLE 4

Understanding Others
Among Three Different School Settings

School setting	Inventory results			
	N	$\bar{X}$	s	s^2
Suburban	33	27.3	.9	.8
Rural	30	26.6	2.1	4.7
Urban	28	27.3	1.67	2.79

TABLE 5

Citizenship Between
Two Different School Settings

School setting	Inventory results			
	N	$\bar{X}$	s	s^2
Rural	30	172.3*	12.5	157.3
Urban	29	185.8*	14.88	221.64

*$p < .01$

TABLE 6

Appreciation of Human Accomplishments
Among Three Different School Settings

School setting	Inventory results			
	N	$\bar{X}$	s	s^2
Suburban	33	117.6*	21.1	446.6
Rural	30	101.6*	15.0	227.7
Urban	29	118.0*	15.42	238.03

*$p < .01$

The results support that hypothesis. Urban *Ss* scored significantly higher than rural *Ss* in Citizenship, and both urban and suburban *Ss* scored significantly higher than rural *Ss* in Appreciation of Human Accomplishments. The means for Understanding Others are in the predicted direction but are not significant.

In a third study, *Ss* individually and in small groups were asked questions relating to specific items on the same three inventories. These questions were quite frequently paraphrases of the original items. When the interviews were conducted individually, the oral responses given by the *Ss* almost always affirmed their written responses. A misunderstanding of the question or an accidental response mark by the *S* was the usual explanation of the discrepancy when it occurred.

As a result of these interviews, the investigators concluded that the *Ss* were in fact responding candidly and openly to the items and not providing the socially expected response. For example, when one male student in the rural class was asked why he indicated that he would not dance with a black girl at a school dance, he said, "Dancing leads to dating, and dating leads to marriage," or words to this effect—presumably implying that interracial marriage was not an appropriate form of behavior in this rural community.

In a fourth study, over 400 5th grade *Ss* in urban, rural and suburban schools were administered three inventories: Self-Esteem, Attitude Toward School and the Assessment of Creative Tendencies. Their 15 teachers separately and independently made judgments about the *Ss*' attitudes toward themselves and toward school, using two scales developed by the EQA staff. Also, on the assumption that curiosity is a component of creative tendencies, an adjective checklist and a curiosity behavior profile developed by Ellen Greenberger of Johns Hopkins University were completed by the teachers. The results revealed that *Ss*' self-reports and teachers' judgments are positively and significantly correlated.

In a fifth study two groups of high school students were administered Creative Potential inventories. Group 1 (N = 1200) was representative of 11th grade students in general from throughout the state. Group 2 (N = 168) represented students who were judged to be creative through selection of their work as outstanding in a regional art contest, a regional or district-wide science fair, or participation in a fine arts program for high school students at Westminster College.

It was hypothesized that Group 2 would score significantly higher in Creative Potential than Group 1. The results support the hypothesis. The Group 2 mean was significantly higher ($p. < .001$) than the Group 1 mean. The results were also practically significant in that only 10 of the 168 *Ss* in Group 2 did not exceed the average of the Group 1 *Ss*. It is interesting to note further that bias would favor no significant differences between the groups, since Group 1 undoubtedly contained a proportion of potentially creative students.

In a sixth study two groups of high school students were administered Creative Output inventories. Group 1 (N = 17,000) was representative of 11th grade students in general from throughout the state. Group 2 was the same group who participated in the fifth study.

It was hypothesized that Group 2 would score significantly higher than Group 1 in Creative Output. The results support the hypothesis. In Creative Output the Group 2 mean was significantly higher than the Group 1 mean ($p < .001$).

The investigators concluded that both the Creative Potential and the Creative Output inventories are valid indicators for differentiating between students in general and those students who have displayed creativity in the fine arts and the sciences.

It may be well at this point to restate the writers' conclusion that validity and reliability are neither completely dichotomous, general nor absolute. Decisions must be made by the users concerning the adequacy with which these concepts are fulfilled in each situation. Judgments must also be made about the adequacy of alternative inventories and the consequences of no measurement at all. Validity studies of each of the EQA inventories continue so that information about the performance of these measures under a variety of circumstances may be gathered in order to assess their appropriateness for any given situation.

References

American Psychological Association, Inc. *Standards for educational and psychological tests and manuals.* Washington, D.C.: APA, 1966.

Coopersmith, S. *The antecedents of self-esteem.* San Francisco: W. H. Freeman and Company, 1967.

Crites, J. O. *The maturity of vocational attitudes in adolescence.* Iowa City: The University of Iowa, 1969.

Cronbach, L. J. *Coefficient alpha and the internal structure of tests. Psychometrika,* 1951, *16* (3), 297-334.

Cronbach, L. J. *Validation of educational measures.* Paper presented at the Invitational Conference on Testing Problems, sponsored by Educational Testing Service, Princeton, 1969.

Kerlinger, F. N. *Foundations of behavioral research.* New York: Holt, Rinehart and Winston, Inc., 1964.

Oppenheim, A. N. *Questionnaire design and attitude measurement.* New York: Basic Books, Inc., 1966.